Assessing Data Augmentation-Induced Bias in Training and Testing of Machine Learning Models

Riddhi More, Jeremy S. Bradbury

Software Engineering & Education Research Lab
Ontario Tech University, Oshawa, Canada

https://www.seerlab.ca

SEER LAB | OntarioTech UNIVERSITY

FAIRNESS WORKSHOP

Data Augmentation

"...involves enhancing the sufficiency and diversity of training examples without explicitly collecting new data... The essence of data augmentation lies in generating new data by altering existing data points through various transformations" [1].

[1] Y. Zhou, C. Guo, X. Wang, Y. Chang, and Y. Wu, "A survey on data augmentation in large model era," arXiv preprint:2401.15422, 2024.

Our analysis of data augmentation is motivated by a desire to understand any potential **bias** considerations.

Data Augmentation in FlakyCat Dataset

- To avoid self-confirmatory bias in our case study, we rely on an existing augmentation technique that was applied to the **FlakyCat** [2] data set by a third-party.
- For each original tests case (**v0**) there are two augmented versions (**v1**, **v2**) created using an adapted Synthetic Minority Over-sampling Technique (SMOTE) that involved generating variants through controlled mutations of non-flakiness-related elements.

[2] A. Akli, G. Haben, S. Habchi, M. Papadakis, and Y. Le Traon, “FlakyCat: Predicting flaky tests categories using few-shot learning,” in Proc.of IEEE/ACM International Conference on Automation of Software Test (AST 2023), May 2023, pp. 140–151.

Dataset

FlakyCat Dataset [1]

- 369 flaky tests – v0
- Augmented versions - v1, v2
- Diverse data from multiple open-source projects
- 5 categories of flakiness root cause.

Class	Data	
	Original	*Augmented*
Async waits	89	285
Unordered collections	45	136
Concurrency	36	113
Time	35	105
Test order dependency	33	99

Table I: Categories of Flakiness in FlakyCat dataset

[1] A. Akli, G. Haben, S. Habchi, M. Papadakis, and Y. Le Traon, “Flakycat: predicting flaky tests categories using few-shot learning,” in 2023 IEEE/ACM Int. Conf. on Automation of Software Test (AST 2023), pp. 140–151.

Research Questions

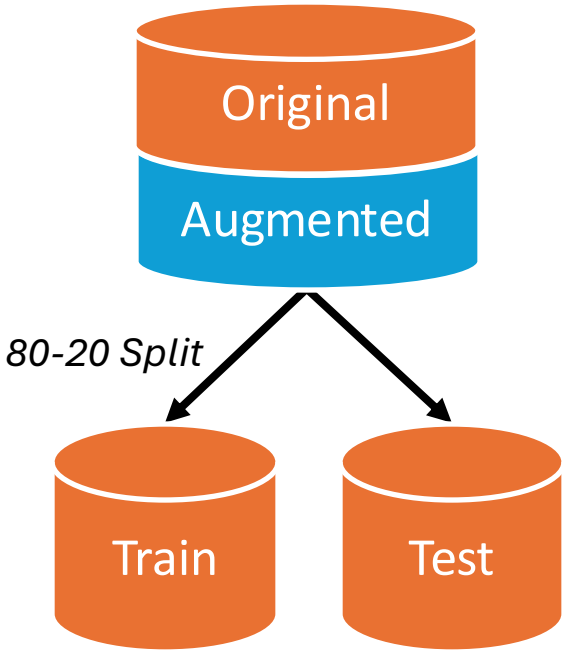
- **RQ1:** What is the **impact** of data augmentation in flaky test classification (FlakyXbert [3]) with the FlakyCat dataset?

- **RQ2:** What is the **bias** of data augmentation in flaky test classification (FlakyXbert) with the FlakyCat dataset?

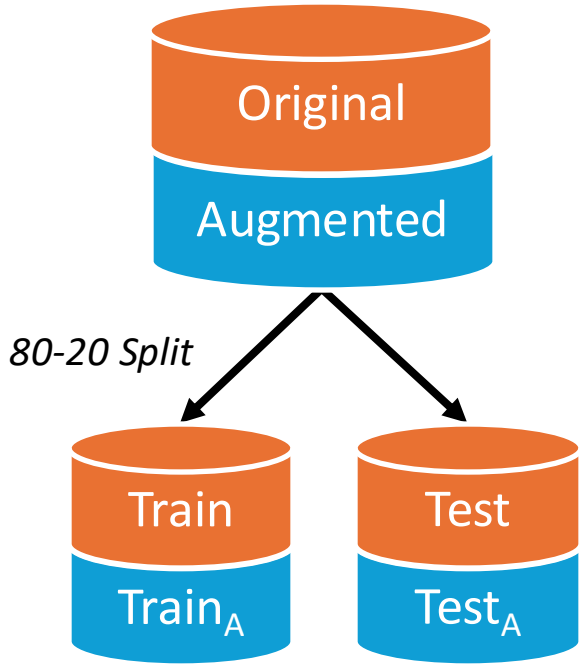
[3] Riddhi More, Jeremy S. Bradbury. “An Analysis of LLM Fine-Tuning and Few-Shot Learning for Flaky Test Detection and Classification.” Proc. of the International Conference on Software Testing, Verification and Validation (ICST 2025), Napoli, Italy, Mar./Apr. 2025.

Experiment 1: Design

- **RQ1:** What is the **impact** of data augmentation in flaky test classification (FlakyXbert) with the FlakyCat dataset?



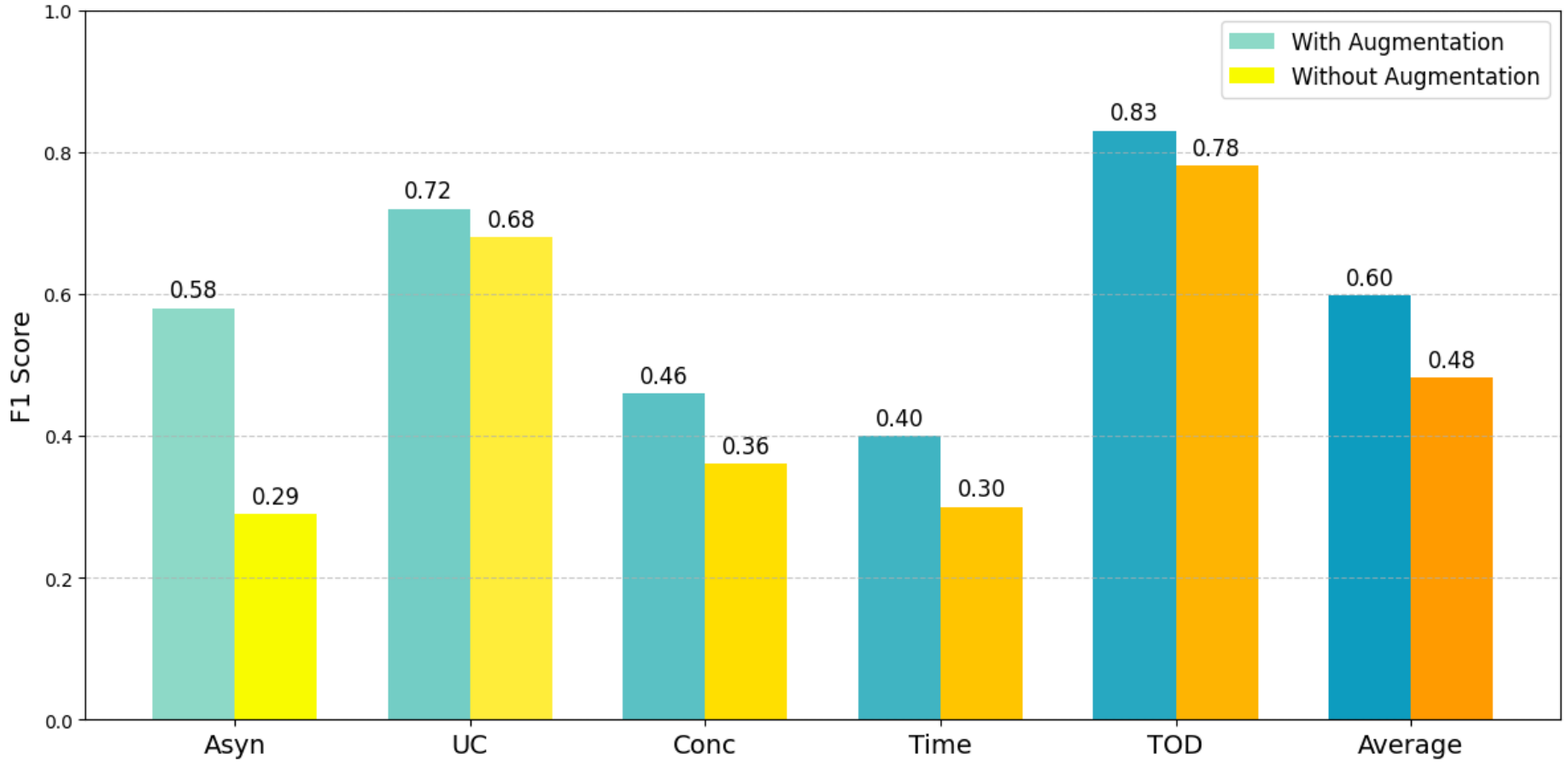
Phase A (Baseline)



Phase B (Augmented)

This ensures a more accurate evaluation of the model's ability to generalize to unseen cases and assess whether it learns meaningful patterns rather than artifacts of augmentation.

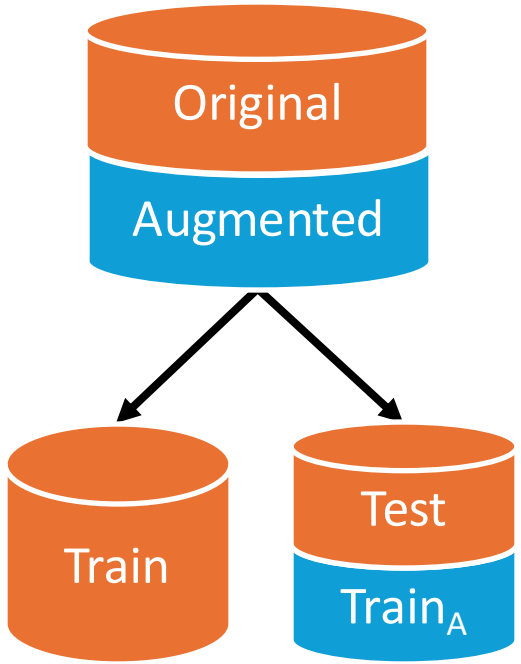
Experiment 1: Results



12% improved F1 score with augmentation.

Experiment 1: Design

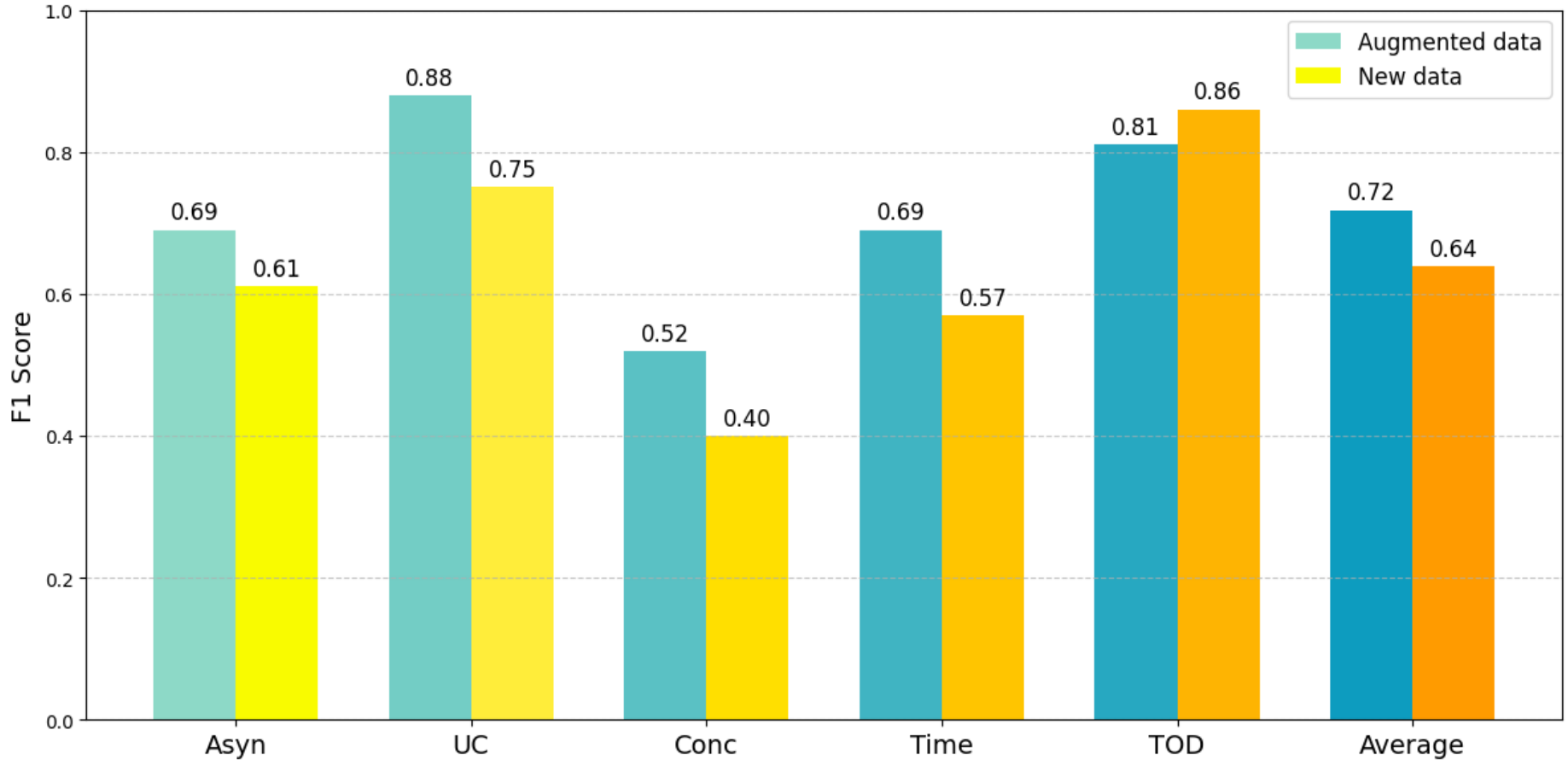
- **RQ2:** What is the **bias** of data augmentation in flaky test classification (FlakyXbert) with the FlakyCat dataset?



Test = new data
Train$_A$ = augmented training data

This design examines whether the model performs differently on augmented variants of familiar cases versus new cases, providing insight into its ability to generalize flaky test characteristics or its tendency to overfit augmentation patterns.

Experiment 2: Results



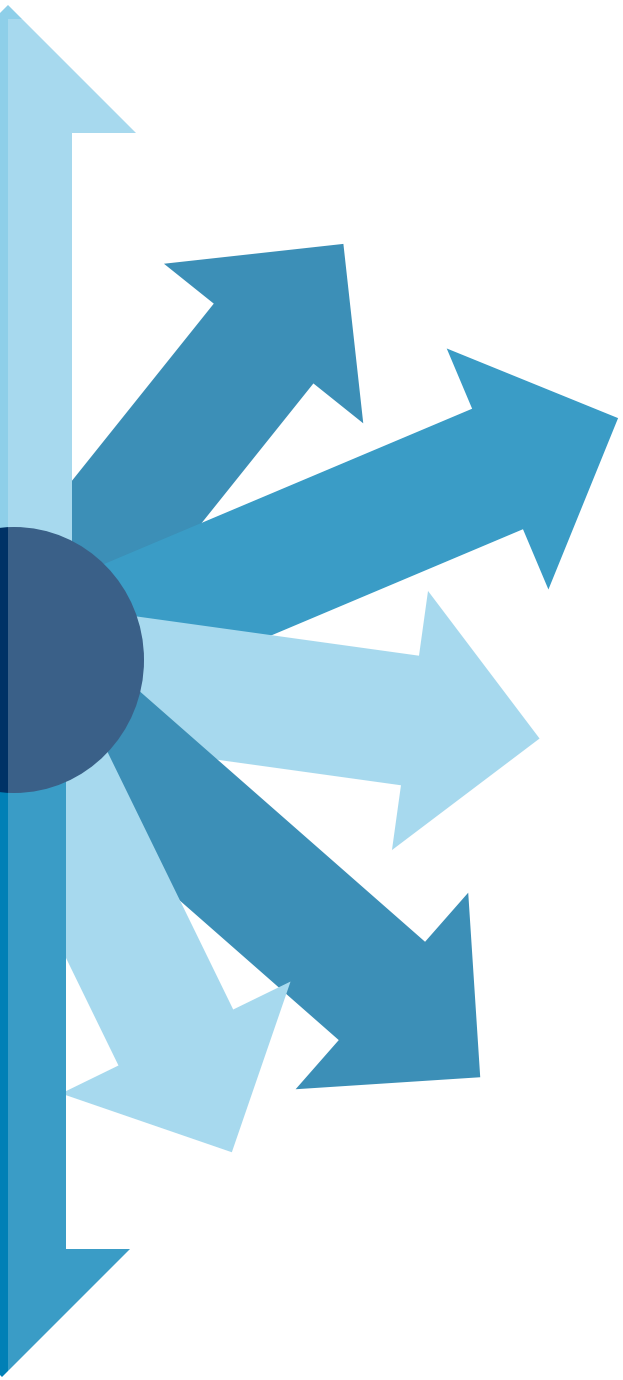
8% improved F1 score with augmented data.

Preliminary Recommendations

1. The systematic performance gap between augmented and new test cases (8% average F1 score difference) suggests **there is a need to maintain a completely separate validation set of original, non-augmented data during model evaluation and testing.**
2. The varying effectiveness of augmentation across different flaky test categories indicates **there is a potential benefit of category-specific augmentation strategies rather than a one-size-fits-all approach to data augmentation.**
3. Bias in data augmentation likely extend beyond flaky test detection. **Researchers in other areas of SE may benefit from assessing bias in augmented data as well as strategies to reduce any identified bias.**

Conclusions

- Our experiments reveal both **benefits** and **biases** from using the augmented FlakyCat dataset
- Our findings highlight the need for **careful use of augmented data** and **robust evaluations to detect biases** and ensure results reflect real-world model performance.



Assessing Data Augmentation-Induced Bias in Training and Testing of Machine Learning Models

Riddhi More, Jeremy S. Bradbury

Software Engineering & Education Research Lab
Ontario Tech University, Oshawa, Canada

https://www.seerlab.ca

SEER LAB | OntarioTech UNIVERSITY

FAIRNESS WORKSHOP